HW 5 Sentimental Classification with Texts

學號：R06323011 系級：經濟碩一 姓名：葉政維

1. (1%) 請說明你實作的 RNN model，其模型架構、訓練過程和準確率為何？

(Collaborators: r05323040 田家駿，討論 Word2Vec 以及疊層的 LSTM)[1]

使用未標籤與有標籤的資料，在 genism 中進行 word2vec 的訓練。相關訓練參數如下圖所示，另外再進行訓練前，我有先將「’ 」去除掉並合併前後文字（如果存在），比如說： can ‘t，就會變成 cant。接著，利用訓練好的 W2V model 去進行將 text 轉換成 a sequence of vector，遇到不知道的文字則去除掉，且長度統一用零向量 pad 成長度 40，最終每一個 text 的維度為（40、100）。最後 RNN 的訓練架構如下圖所示（註：兩次 dropout rate 分別為 0.4、0.2；loss func=binary cross-entropy；opt=adam；epoch=7）。最終模型準確率為 0.8261（private score）

```
t0 = time()
model = Word2Vec(X_wseq, sg=0, size=100, window=5, min_count=3)
print ('Time Consumption:', time() - t0)
```

Layer (type)	Output Shape	Param #
lstm_3 (LSTM)	(None, 40, 200)	240800
dropout_3 (Dropout)	(None, 40, 200)	0
lstm_4 (LSTM)	(None, 100)	120400
dense_4 (Dense)	(None, 100)	10100
dropout_4 (Dropout)	(None, 100)	0
dense_5 (Dense)	(None, 20)	2020
dense_6 (Dense)	(None, 2)	42

Total params: 373,362
Trainable params: 373,362
Non-trainable params: 0

[1] 參考網站：http://kavita-ganesan.com/gensim-word2vec-tutorial-starter-code/#.WwKkM1OFN-U

2. **(1%) 請說明你實作的 BOW model，其模型架構、訓練過程和準確率為何？**

   本題只利用到標籤資料，首先進行 tokenize（字庫上限為 20000），並轉換成 text matrix，其值透過 TFIDF 產生。最後將 text matrix 餵進 DNN，其架構如下（註：其他設定同題 1）。此模型準確率為 0.7898（private score）。

Layer (type)	Output Shape	Param #
dense_38 (Dense)	(None, 500)	10000500
dropout_21 (Dropout)	(None, 500)	0
dense_39 (Dense)	(None, 50)	25050
dropout_22 (Dropout)	(None, 50)	0
dense_40 (Dense)	(None, 2)	102

Total params: 10,025,652
Trainable params: 10,025,652
Non-trainable params: 0

3. **(1%) 請比較 bag of word 與 RNN 兩種不同 model 對於"today is a good day, but it is hot"與"today is hot, but it is a good day"這兩句的情緒分數，並討論造成差異的原因。**

   本題使用題 1、2 的模型進行預測。RNN 的負向情緒分數(class 0)分別為 0.9497、0.0135，而 BOW 的分數則同樣是 0.3924。這是因為 BOW 並不考慮文字的順序，此外，從分數的絕對大小看來，RNN 比起 BOW 更肯定情緒類別為何，顯示文字順序（尤其在轉折詞 but 前後）是很有價值的資訊。

4. **(1%) 請比較"有無"包含標點符號兩種不同 tokenize 的方式，並討論兩者對準確率的影響。**

   題 1 的結果即有包含標點符號（只有用到 , . ! ? : ; ）。在同一個（和題 1 不同）RNN 訓練架構下，有無使用標點符號的結果分別為 0.8241 與 0.8165。此差異的原因是因為部分標點符號確實能反映情緒，甚至同一個文字搭配不同的標點符號情緒也可能截然不同。

5. **(1%) 請描述在你的 semi-supervised 方法是如何標記 label，並比較有無 semi-surpervised training 對準確率的影響。**[2]

本題建立在題 1 的模型基礎之上，進行半監督訓練。訓練流程是先用原模型對未標籤資料進行預測，如某一類別的準確率高達 87%，則給予對應的標籤、納入有標籤資料並在下一輪一起進行訓練。本人的配備與耐心有限，因此事實上每一輪只取約 40 萬筆的未標籤資料去預測，並透過 data_generator 餵資料的方式，減少記憶體負擔。最終模型進行了 2 輪，每輪跑 3 個 epochs，最後依驗證準確率取過程中的最佳模型。此模型準確率為 0.8290，比起題 1 有所進步，

[2] 參考資料 https://stanford.edu/~shervine/blog/keras-how-to-generate-data-on-the-fly.html